

CHAPTER 9

Statistics and Probability with Python

In the previous chapter, we learned about how to apply your knowledge of data analysis by solving some case studies.

Now, in the final part of this book, we learn about essential concepts in statistics and probability and understand how to solve statistical problems with Python. The topics that we cover include permutations and combinations, probability, rules of probability and Bayes theorem, probability distributions, measures of central tendency, dispersion, skewness and kurtosis, sampling, central limit theorem, and hypothesis testing. We also look at confidence levels, level of significance, p-value, hypothesis testing, parametric tests (one- and two-sample z-tests, one- and two-sample t-tests, paired tests, analysis of variance [ANOVA]), and nonparametric tests (chi-square test).

Permutations and combinations

Let us look at a few definitions, formulae, and examples that will help us understand the concepts of permutations and combinations.

Combinations: The various ways in which we can select a group of objects.

The following formula gives the number of combinations we can form from a given number of objects:

$$nC_r = \frac{n!}{r!(n-r)!}$$

In the preceding formula, n is the total number of objects in the set from which a smaller subset of objects is drawn, c is the number of combinations, and r is the number of objects in the subset. The exclamation mark symbol (!) denotes the factorial of a number. For example, $x!$ is the product of all integers from 1 up to and including x .

$$(x! = x * (x-1) * (x-2) \dots * 1)$$

Let us now solve a simple question involving combinations.

Question: Find the number of ways in which an ice cream sundae containing three flavors can be created out of a total of five flavors.

Answer: Let the five flavors be A, B, C, D, and E. Working out this problem manually, the following combinations can be obtained:

A, B, C | B, C, D | A, C, D | A, B, D | C, D, E | B, D, E | A, B, E | A, D, E | A, C, E | B, C, E

There are ten combinations, as we can see. If we apply the nc_r formula, where n is 5 and r is 3, we get the same answer ($5C_3 = 10$).

Let us now look at what permutations are.

Permutations are similar to combinations, but here, the order in which the objects are arranged matters.

The following formula gives the number of permutations:

$$n_{P_r} = \frac{n!}{(n-r)!}$$

Considering the same ice cream example, let us see how many permutations we can obtain, that is, the number of ways in which three flavors can be selected and arranged out of a total of five flavors.

1. ABC | CBA | BCA | ACB | CAB | BAC
2. BCD | CDB | BDC | CBD | DBC | DCB
3. ACD | ADC | DAC | DCA | CAD | CDA
4. ABD | ADB | BAD | BDA | DAB | DBA
5. CDE | CED | DCE | DEC | ECD | EDC

6. BDE|BED|DBE|DEB|EBD|EDB
7. ABE|AEB|BEA|BAE|EAB|EBA
8. ADE|AED|DAE|DEA|EAD|EDA
9. ACE|AEC|CAE|CEA|EAC|ECA
10. BCE|BEC|CBE|CEB|EBC|ECB

As we can see, we can get six possible arrangements for each combination. Hence, the total number of arrangements = $10 \times 6 = 60$. The formula nPr (where $n=5$ and $r=3$) also gives the same answer (60).

Another approach to solving questions involving permutations is as follows:

1. First, select the items: Select three flavors from five in 5C_3 ways
2. Now arrange the three items in $3!$ ways
3. Multiply the results obtained in step 1 and step 2. Total number of permutations = ${}^5C_3 \times 3! = 60$

Now that we have understood the concepts of permutations and combinations, let us look at the essentials of probability.

Probability

Given below are a few important concepts related to probability.

Random experiment: This is any procedure that leads to a defined outcome and can be repeated any number of times under the same conditions.

Outcome: The result of a single experiment.

Sample space: Exhaustive list containing all possible outcomes of an experiment.

Event: An event can comprise a single outcome or a combination of outcomes. An event is a subset of the sample space.

Probability: A quantitative measure of the likelihood of the event. The probability of any event always lies between 0 and 1. 0 denotes that the event is not possible, while 1 indicates that the event is certain to occur.

If the letter X denotes our event, then the probability is given by the notation $P(X) = \frac{N(X)}{N(S)}$

Where $N(X)$ =number of outcomes in event x

$N(S)$ = total number of outcomes in the sample space

Solved example: Probability

The following is a simple probability question.

Question: In an experiment, a die is rolled twice. Find the probability that the numbers obtained in the two throws add up to 10.

Solution:

Event A: The first die is rolled.

Event B: The second die is rolled.

Sample space: A die contains the numbers 1 to 6, which are equally likely to appear. The total number of outcomes, when one die is rolled, is six. Since events “A” and “B” are independent, the total number of outcomes for both the events = $6*6 = 36$.

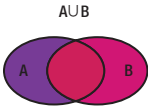
Event X: The sum of the two numbers is 10. The possible outcomes that lead to this result include {4,6}, {6,4}, and {5,5}; that is, three possible outcomes lead to a sum of 10.

$P(X)$ =Probability of obtaining a sum of 10=Number of outcomes in event X/Total Sample Space= $3/36=0.0833$

Rules of probability

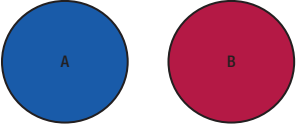
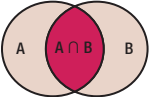
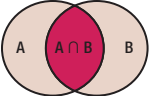
Let us understand the various rules of probability, explained in Table 9-1.

Table 9-1. Rules of Probability

Rule	Description	Formula	Venn diagram
Addition rule	The addition rule determines the probability of either of two events occurring.	$P(A \cup B) = P(A)+P(B)-P(A \cap B)$	

(continued)

Table 9-1. (continued)

Rule	Description	Formula	Venn diagram
Special rule of addition	This rule applies to mutually exclusive events. Mutually exclusive events are those that cannot co-occur. For mutually exclusive events, the probability of either of the events occurring is simply the sum of the probability of each of the events.	$P(A \cup B) = P(A) + P(B)$	<p>MUTUALLY EXCLUSIVE</p> 
Multiplication rule	The multiplication rule is a general rule that provides the probability of two events occurring together.	$P(A \cap B) = P(A) * P(B/A)$ $P(B/A)$ is the <i>conditional probability</i> of event B happening given that event A has already occurred.	
Special rule of multiplication	This rule applies to independent events. For independent events, the probability of the events occurring together is simply the product of probabilities of the events.	$P(A \cap B) = P(A) * P(B)$	

Note that the formulae listed in Table 9-1 provide the rules for two events, but these can be extended to any number of events.

Conditional probability

Conditional probability involves calculating the probability of an event, after taking into consideration the probability of another event that has already occurred. Consider Figure 9-1, which illustrates the principle of conditional probability.

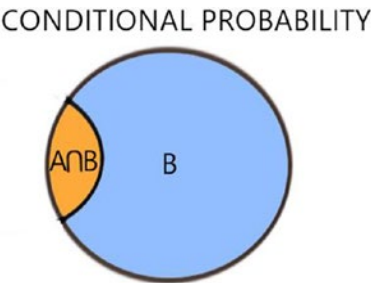


Figure 9-1. *Conditional probability*

Figure 9-1 shows that if the event “A” is dependent on event “B,” then the sample space is event “B” itself. For example, let A be the event that a customer purchases a product from an online retailer. Let the probability of the event be 0.5., or in other words, $P(A)=0.5$.

Now, let B be the event in which the product that the customer intends to purchase has received a negative review. The probability of the customer buying the product may now be less than what it was earlier due to this negative review. Let us say that now there is only a 30% chance that they purchase the product. In other words, $P(A/B)$ = probability of the customer buying a product given that it has received a negative review = 0.3.

The formula for conditional probability is $P(A/B) = P(A \cap B)/P(B)$.

Bayes theorem

Bayes theorem is a theorem used to calculate the conditional probability of an event, given some evidence related to the event. Bayes theorem establishes a mathematical relationship between the probability of an event and prior knowledge of conditions related to it. As evidence related to an event accumulates, the probability of this event can be determined more accurately.

Questions involving Bayes theorem are different from conventional probability questions, where we generally calculate the probability of events that may occur in the future. For instance, in a simple conditional probability question, we might be asked to calculate the probability of a person getting diabetes, given that they are obese. In Bayes theorem, we go backward and calculate the probability of a person being obese, given that they have diabetes. That is, if a person tested positive for diabetes, Bayes theorem tests the hypothesis that he is obese. The formula for Bayes theorem is as follows:

$$P(A/B) = \frac{P(B/A) * P(A)}{P(B)}$$

$P(A/B)$, also known as the posterior probability, is what we want to compute, that is, the probability of the hypothesis being true, given the data on hand.

$P(B/A)$ is the probability of obtaining the evidence, given the hypothesis.

$P(A)$ is the *prior probability*, that is, the probability of the hypothesis being true before we have any data related to it.

$P(B)$ is the general probability of occurrence of the evidence, without any hypothesis, also called the *normalizing constant*.

Applications of Bayes theorem: Given below are a few areas where Bayes theorem can be applied.

- Medical diagnostics
- Finance and risk assessment
- Email spam classification
- Law enforcement and crime detection
- Gambling

Now, let us understand the practical application of Bayes theorem in a few of these areas using a couple of examples.

Application of Bayes theorem in medical diagnostics

Bayes theorem has applications in the field of medical diagnostics, which can be understood with the help of the following hypothetical example.

Question: Consider a scenario where a person has tested positive for an illness. This illness is known to impact about 1.2% of the population at any given time. The diagnostic test is known to have an accuracy of 85%, for people who have the illness. The test accuracy is 97% for people who do not have the illness. What is the probability that this person suffers from this illness, given that they have tested positive for it?

Solution:

Assessing the accuracy of medical tests is one of the applications of Bayes theorem.

Let us first define the events:

A: The person has the illness, also called the hypothesis.

$\sim A$: The person does not have the illness.

B: The person has tested positive for the illness, also called the evidence.

$P(A/B)$: Probability that this person has the illness given that they have tested positive for it, also called the *posterior probability* (which is what we need to calculate).

$P(B/A)$: Probability that the person has tested positive for it given that they have the illness. This value is 0.85 (as given in the question that this test has 85% accuracy for people who suffer from the illness).

$P(A)$: *Prior probability* or the probability that the person has the illness, without any evidence (like a medical test). This value is 0.012 (as given in the question that this illness affects 1.2% of the population).

$P(B)$: Probability that this person has tested positive for this illness. This probability can be calculated in the following manner.

There are two ways this person could test positive for this illness:

- They have the illness and have tested positive (true POSITIVE) - the probability of this occurring can be calculated as follows:

$$P(B/A)*P(A)=0.85*0.012=0.0102.$$

- They do not have the illness, but the test was inaccurate, and they have tested positive for it (false positive) – the probability of this occurring can be calculated as follows:

$$P(B/\sim A)*P(\sim A)=(1-0.97)*(1-0.012)=0.02964.$$

Here, $P(B/\sim A)$ refers to the probability that the test was positive for a person who did not have the illness, that is, the probability that the test was inaccurate for the person who does not suffer from the illness.

Since this test is 97% accurate for people who do not have this illness, it is inaccurate for 3% of the cases.

In other words, $P(B/\sim A) = 1 - 0.97$.

Similarly, $P(\sim A)$ refers to the probability that the person does not have the illness. Since we have been provided the data that the incidence of this illness is 1.2%, $P(\sim A)$ is $= 1 - 0.012$.

$P(B)$, the denominator in the Bayes theorem equation, is the union of the preceding two probabilities $= (P(B/A)*P(A)) + (P(B/\sim A)*P(\sim A)) = 0.0102 + 0.2964 = 0.03984$

We can now calculate our final answer by plugging in the values for the numerator and denominator in the Bayes theorem formula.

$$P(A/B) = P(B/A) * P(A) / P(B) = 0.85 * 0.012 / 0.03984 = 0.256$$

Using the Bayes theorem, we can now conclude that even with a positive medical test, this person only has a 25.6% chance of suffering from this illness.

Another application of Bayes theorem: Email spam classification

Let us look at another application of Bayes theorem in the area of email spam classification. Before the era of spam filters, there was no way of separating unsolicited emails from legitimate ones. People had to sift through their emails to identify spam manually. Nowadays, email spam filters have automated this task and are quite efficient at identifying spam emails and keeping only ham (nonspam) emails in the box. The Bayesian approach forms the principle behind many spam filters. Consider the following example:

Question: What is the probability of a mail being spam, given that it contains the word “offer”? Available data indicates that 50% of all emails are spam mails. 9% of spam emails contain the word “offer,” and 0.4 % of ham emails contain the word “offer.”

Answer:

Defining the events and probabilities as follows:

A: Email is “spam”

~A: Email is “ham”

B: Email contains the word “offer”

$P(A) = 0.5$ (assuming 50% of emails are spam mails)

$P(B/A) =$ Probability of spam mail containing the word “offer” = 0.09 (9%)

$P(B/\sim A) =$ Probability of ham mail containing the word “offer” = 0.004 (0.4%)

Applying the Bayes theorem:

$P(A/B) = (0.09*0.5)/(0.09*0.5)+(0.004)*(0.5) = 0.957$

In other words, the probability of the mail being a spam mail given that it has the word “offer” is 0.957.

SciPy library

Scipy, a library for mathematical and scientific computations, contains several functions and algorithms for a wide range of domains, including image processing, signal processing, clustering, calculus, matrices, and statistics. Each of these areas has a separate submodule in SciPy. We use the *scipy.stats* submodule in this chapter, and apply the functions from this submodule for statistical tests and different types of distributions. This module also contains functions for distance calculations, correlations, and contingency tables.

Further reading:

Read more about the *scipy.stats* module and its functions:

<https://docs.scipy.org/doc/scipy/reference/stats.html>

Probability distributions

To understand probability distributions, let us first look at the concept of random variables, which are used to model probability distributions.

Random variable: A variable whose values equal the numeric values associated with the outcomes of a random experiment.

Random variables are of two types:

1. Discrete random variables can take a finite, countable number of values. For example, a random variable for the Likert scale, used for surveys and questionnaires to assess responses, can take values like 1, 2, 3, 4, and 5. The *probability mass function*, or *PMF*, associated with a discrete random variable is a function that provides the probability that this variable is exactly equal to a certain discrete value.
2. Continuous random variables can take infinitely many values. Examples include temperature, height, and weight. For a continuous variable, we cannot find the absolute probability. Hence, we use the *probability density function*, or *PDF*, for continuous variables (the equivalent of PMF for discrete variables). The PDF is the probability that the value of a continuous random variable falls within a range of values.

The cumulative distribution function (CDF) gives the probability of a random variable being less than or equal to a given value. It is the integral of the PDF and gives the area under the curve defined by the PDF up to a certain point.

In the following section, we cover the two types of probability distributions for discrete random variables: binomial and Poisson.

Binomial distribution

In a binomial experiment, there are several independent trials, with every trial having only *two possible outcomes*. These outcomes are the two values of the binomial discrete random variable. A basic example of a binomial distribution is the repeated toss of a coin. Each toss results in only two outcomes: Heads or Tails.

The following are the characteristics of a binomial distribution:

1. There are n identical trials
2. Each trial results in either one of only two possible outcomes
3. The outcomes of one trial do not affect the outcomes of other trials
4. The probability of success (p) and failure (q) is the same for each trial
5. The random variable represents the number of successes in these n trials and can at most be equal to n
6. The mean and variance of the binomial distribution are as follows:

Mean = $n * p$ (number of trials * probability of success)

Variance = $n * p * q$ (number of trials * probability of success * probability of failure)

The PMF, or the probability of r successes in n attempts of an experiment, is given by the following equation:

$$P(X=r) = {}^n C_r p^r q^{n-r}$$

Where p is the probability of success, q is the probability of failure, and n is the number of trials

The shape of a binomial distribution

The binomial distribution resembles a skewed distribution, but it approaches symmetry and looks like a normal curve as n increases and p becomes smaller, as shown in Figure 9-2.

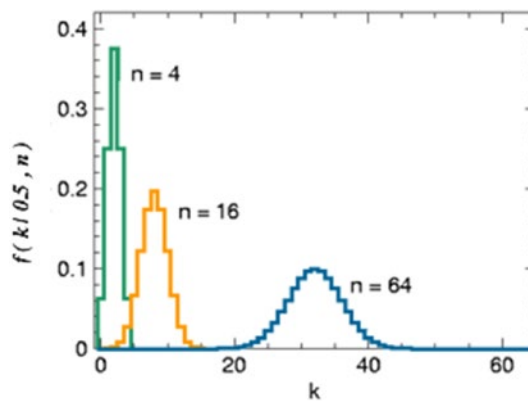


Figure 9-2. Binomial distribution for different values

Question: The metro rail company surveys eight senior citizens traveling in a subway train about their satisfaction with the new safety features introduced in the subway trains. Each response has only two values: yes or no. Let us assume that the probability of a “yes” response is 0.6, and the probability of a “no” response is 0.4 based on historical survey.

Calculate the probability that

1. Exactly three people are satisfied with the metro’s new safety features
2. Fewer than five people are satisfied

Solution:

1. For part 1 of the question: We can either use the formula ${}^nC_r p^r q^{n-r}$ or solve it using a Scipy function (*stats.binom.pmf*), as shown in the following:

CODE:

```
import scipy.stats as stats
n,r,p=8,3,0.6
stats.binom.pmf(r,n,p)
```

Output:

```
0.12386304000000009
```

Explanation: First, the *scipy.stats* module needs to be imported. Then we define three variables - n (the number of trials), r (the number of successes), and p (the probability of failure). After this, the PMF for binomial distributions (*stats.binom.pmf*) is called, and we pass three parameters - r , n , and p in that order. The *pmf* function is used since we are calculating the probability of a discrete variable.

2. For part two of the question: Since we need to calculate the probability that fewer than five people are satisfied, the limiting value of the variable is 4.

The following equation gives the probability we need to calculate:

$$P(X \leq 4) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4)$$

We can either apply the formula ${}^nC_r p^r q^{n-r}$ to calculate the values for $r = 0, 1, 2, 3$, and 4 or solve using the *stats.binom.cdf* function in Scipy as follows:

CODE:

```
import scipy.stats as stats
n,r,p=8,4,0.6
stats.binom.cdf(r,n,p)
```

Output:

```
0.40591359999999976
```

Explanation: We use the CDF function when we calculate the probability for more than one value of x .

Poisson distribution

A Poisson distribution is a distribution that models the number of events that occur over a given interval (usually of time, but can also be an interval of distance, area, or volume). The average rate of occurrence of events needs to be known.

The PMF for a Poisson distribution is given by the following equation:

$$P(x=r) = \frac{\lambda^r e^{-\lambda}}{r!}$$

where $P(x=r)$ is the probability of the event occurring r number of times, r is the number of occurrences of the event, and λ^r represents the average/expected number of occurrences of that event.

The Poisson distribution can be used to calculate the number of occurrences that occur over a given period, for instance:

- number of arrivals at a restaurant per hour
- number of work-related accidents occurring at a factory over a year
- number of customer complaints at a call center in a week

Properties of a Poisson distribution:

1. Mean=variance= λ . In a Poisson distribution, the mean and variance have the same numeric values.
2. The events are independent, random, and cannot occur at the same time.
3. When n is >20 and p is <0.1 , a Poisson distribution can approximate the binomial distribution. Here, we substitute $\lambda = np$.
4. When the value of n is large, p is around 0.5, and $np > 0.5$, a normal distribution can be used to approximate a binomial distribution.

The shape of a Poisson distribution

A Poisson distribution is skewed in shape but starts resembling a normal distribution as the mean (λ) increases, as shown in Figure 9-3.

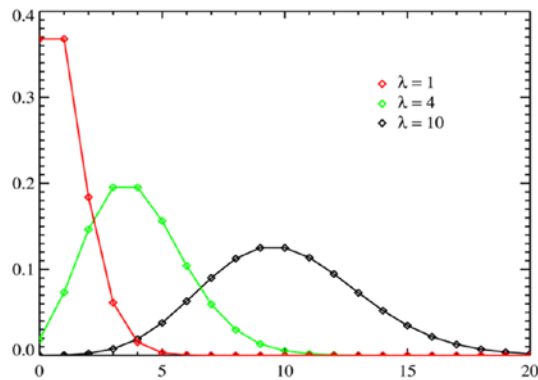


Figure 9-3. Poisson distribution

Solved example for the Poisson distribution:

In a subway station, the average number of ticket-vending machines out of operation is two. Assuming that the number of machines out of operation follows a Poisson distribution, calculate the probability that a given point in time:

1. Exactly three machines are out of operation
2. More than two machines are out of operation

Solution:

1. We can either use the formula $\frac{\lambda^r e^{-\lambda}}{r!}$ or solve it in Python as follows:

CODE:

```
import scipy.stats as stats
λ=2
r=3
stats.poisson.pmf(r,λ)
```

Output:

```
0.18044704431548356
```

Explanation: First, the *scipy.stats* module needs to be imported. Then we define two variables - λ (the average) and r (the number of occurrences of the event). Then, the PMF for a Poisson distribution (*stats.poisson.pmf*) is called, and we pass the two arguments to this function, r and λ , in that order.

2. Since we need to calculate the probability that more than two machines are out of order, we need to calculate the following probability:

$$P(x > 2), \text{ or } (1 - p(x=0) - p(x=1) - p(x=2)).$$

This can be computed using the *stats.poisson.cdf* function, with $r=2$.

CODE:

```
import scipy.stats as stats
λ=2
r=2
1-stats.poisson.cdf(r,λ)
```

Output:

```
0.3233235838169366
```

Explanation: We follow a similar method as we did for the first part of the question but use the CDF function (*stats.poisson.cdf*) instead of PMF (*stats.poisson.pmf*).

Continuous probability distributions

There are several continuous probability distributions, including the normal distribution, Student's T distribution, the chi-square, and ANOVA distribution. In the following section, we explore the normal distribution.

Normal distribution

A normal distribution is a symmetrical bell-shaped curve, defined by its mean (μ) and standard deviation (σ), as shown in Figure 9-4.

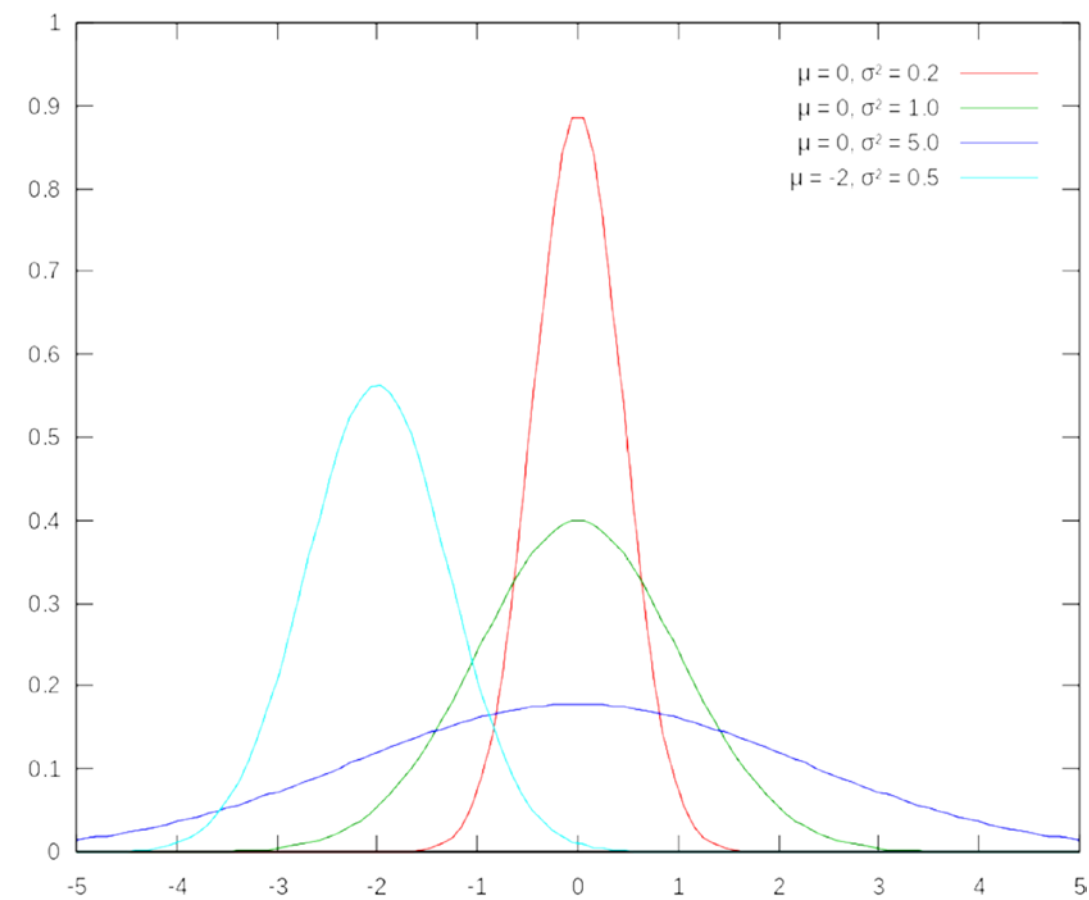


Figure 9-4. Normal distribution

All the four curves in Figure 9-4 are normal distributions. The mean is represented by the symbol μ (mu) and the standard deviation by the symbol σ (sigma)

Characteristics of a normal distribution

1. The central value (μ) is also the mode and the median for a normal distribution
2. Checking for normality: In a normal distribution, the difference between the 75th percentile value (Q3) and the 50th percentile value (median or Q2) equals the difference between the median (Q2) and the 25th percentile. In other words,

$$Q_3 - Q_2 = Q_2 - Q_1$$

If the distribution is skewed, this equation does not hold.

In a right-skewed distribution, $(Q_3 - Q_2) > (Q_2 - Q_1)$

In a left-skewed distribution, $(Q_2 - Q_1) > (Q_3 - Q_2)$

Standard normal distribution

To standardize units and compare distributions with different means and variances, we use a standard normal distribution.

Properties of a standard normal distribution:

- The standard normal distribution is a normal distribution with a mean value of 0 and a standard deviation as 1.
- Any normal distribution can be converted into standard normal distribution using the following formula:

$$z = \frac{(x - \mu)}{\sigma}$$
 where μ and σ are the mean and variance of the original normal distribution.
- In a standard normal distribution,
 - 68.2% of the values lie within 1 standard deviation of the mean
 - 95.4% of the values lie between 2 standard deviations of the mean
 - 99.8% lie within 3 standard deviations of the mean

This distribution of values is shown in Figure 9-5.

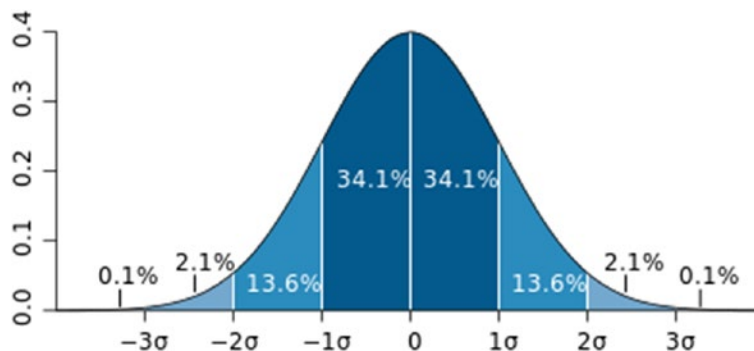


Figure 9-5. Standard normal distribution

- The area under the standard normal distribution between any two points represents the proportion of values that lies between these two points. For instance, the area under the curve on either side of the mean is 0.5. Put in another way, 50% of the values lie on either side of the mean.

There are two types of questions involving normal distributions:

1. Calculating probability/proportion corresponding to the value of a variable: The z-value is calculated using the formula $z = \frac{(x - \mu)}{\sigma}$, and this z-value is then passed as an argument to the *stats.norm.cdf* function
2. Calculating the value of the variable corresponding to a given probability: First, the z-value is obtained by passing the probability value as an argument to the *stats.norm.ppf* function. Then, we obtain the value of the variable (x) corresponding to the z-value by substituting values in the following formula: $z = \frac{(x - \mu)}{\sigma}$

Solved examples: Standard normal distribution

Question: An IT team in a software company is inspecting some laptops. The team needs to select the top 1% of the laptops, with the criterion being the fastest boot times. The average boot time is 7 seconds, with a standard deviation of 0.5 seconds. What would be the cutoff boot time necessary for selection?

Solution:

Step 1: Since the criterion is fast boot time, the boot times of interest lie on the lower left end of the distribution, as shown in Figure 9-6.



Figure 9-6. Lower-tail test (standard normal distribution)

The area of the curve to the right of this tail is 0.99. We calculate the z-value, corresponding to a probability value of 0.99, using the *stats.norm.ppf* function:

CODE:

```
stats.norm.ppf(0.99)
```

Output:

```
2.3263478740408408
```

Since this is a lower-tail test, we take the value of z as -2.33 (this value is negative as it lies to the left of the mean). We can also verify this using the z-table by calculating the z-value corresponding to a probability of 0.99.

Step 2: Apply the following formula and calculate x

$$z = (x - \mu) / \sigma$$

where $z = -2.33$, $\mu = 7$, and $\sigma = 0.5$. We need to calculate the value of x:

CODE:

```
x = (-2.33 * 0.5) + 7
```

Output: 5.835

Inference: The required boot time is 5.835 seconds

Example 2 (standard normal distribution):

A company manufactures tube lights, where the life (in hours) of these tube lights follows a normal distribution with a mean of 900 (hrs) and a standard deviation of 150 (hrs). Calculate the following:

- (1) The proportion of tube lights that fail within the first 750 hours
- (2) The proportion of tube lights that fail between 800 and 1100 hours
- (3) After how many hours would 20% of the tube lights fail?

Solution for Example 2 (standard normal distribution):

- (1) Calculate the z-value corresponding to $X=750$ and obtain the corresponding probability:

CODE:

```
x=750
μ=900
σ=150
z=(x-μ)/σ
z
stats.norm.cdf(z)
```

Output:

```
0.15865525393145707
```

Inference: 15.8% of the tube lights fail within the first 750 hours.

- (2) Calculate the z-values corresponding to x-values of 800 and 1100, respectively, and subtract the probabilities corresponding to these z-values.

CODE:

```
x1=800
x2=1100
μ=900
σ=150
z1=(x1-μ)/σ
z2=(x2-μ)/σ
p2=stats.norm.cdf(z2)
p1=stats.norm.cdf(z1)
p2-p1
```

Output:

```
0.6562962427272092
```

Inference: Around 65.6% of the tube lights, with a lifetime between 800 and 1100 hours, fail.

- (3) Calculate the z-value corresponding to a probability of 0.2 and calculate x by substituting z, μ , and σ in the formula $z = \frac{(x - \mu)}{\sigma}$

CODE:

```
z=stats.norm.ppf(0.2)
μ=900
σ=150
x=μ+σ*z
x
```

Output:

```
773.7568149640629
```

Inference: After a lifetime of around 774 hours, 20% of the tube lights fail.

Measures of central tendency

The central tendency is a measure of the central value among a set of values in a dataset. The following are some of the measures of central tendency:

Mean: This is the average of values in a dataset.

Median: This is the middle number when the values in the dataset are arranged size-wise.

Mode: The most frequently occurring value in a dataset with discrete values.

Percentile: A percentile is a measure of the percentage of values below a particular value. The median corresponds to the 50th percentile.

Quartile: A quartile is a value that divides the values in an ordered dataset into four equal groups. Q1 (or the first quartile) corresponds to the 25th percentile, Q2 corresponds to the median, and Q3 corresponds to the 75th percentile.

Measures of dispersion

The measures of dispersion give a quantitative measure of the spread of a distribution. They provide an idea of whether the values in a distribution are situated around the central value or spread out. The following are the commonly used measures of dispersion.

Range: The range is a measure of the difference between the lowest and highest values in a dataset.

Interquartile range: A measure of the difference between the third quartile and the first quartile. This measure is less affected by extreme values since it focuses on the values lying in the middle. The interquartile range is a good measure for skewed distributions that have outliers. The interquartile range is denoted by $IQR = Q3 - Q1$.

Variance: This is a measure of how much values in a dataset are scattered around the mean value. The value of the variance is a good indication of whether the mean is representative of values in the dataset. A small variance would indicate that the mean is an appropriate measure of central tendency. The following formula gives the variance:

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N},$$

Where μ is the mean, and N is the number of values in the dataset.

Standard deviation: This measure is calculated by taking the square root of the variance. The variance is not in the same units as the data since it takes the square of the differences; hence taking the square root of the variance brings it to the same units as the data. For instance, in a dataset about the average rainfall in centimeters, the variance would give the value in cm^2 , which would not be interpretable, while the standard deviation in cm would give an idea of the average rainfall deviation in centimeters.

Measures of shape

Skewness: This measures the degree of asymmetry of a distribution, as shown in Figure 9-7.

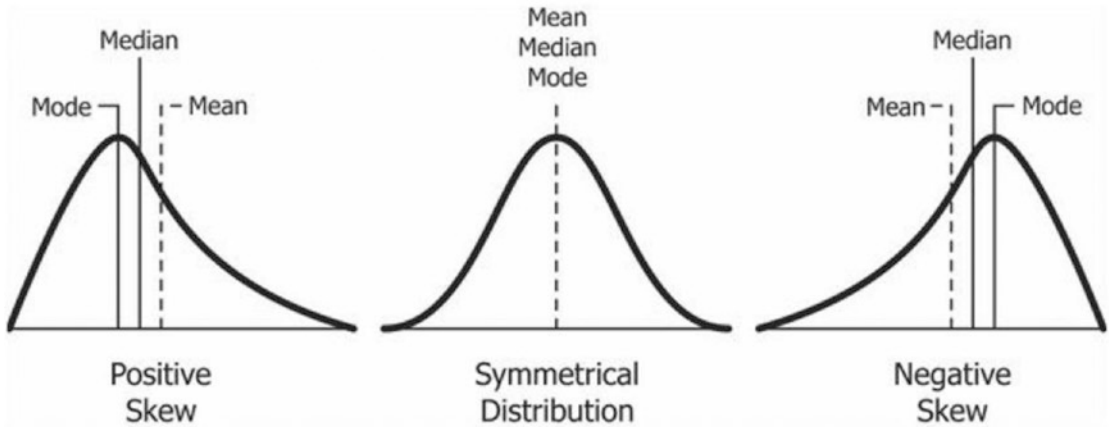


Figure 9-7. *Distributions with varied skewness*

We can observe the following from the Figure 9-7:

In a positively skewed distribution, $\text{mean} > \text{median}$

In a negatively skewed distribution, $\text{mean} < \text{median}$

In a perfectly symmetrical distribution, $\text{mean} = \text{median} = \text{mode}$

Kurtosis

Kurtosis is a measure of whether a given distribution of data is curved, peaked, or flat.

A mesokurtic distribution has a bell-shaped curve. A leptokurtic distribution is one with a marked peak. A platykurtic distribution, as the name indicates, has a flat curve. These distributions are shown in Figure 9-8.

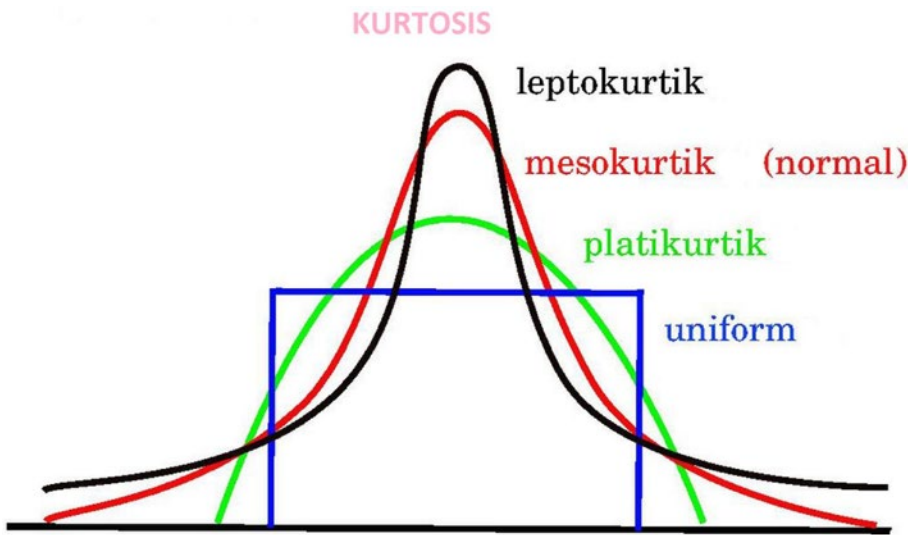


Figure 9-8. Representation of kurtosis

Solved Example:

The weight of children (in kgs) aged 3-7 in a primary school is as follows: 19, 23, 19, 18, 25, 16, 17, 19, 15, 23, 21, 23, 21, 11, 6. Let us calculate the measures of central tendency, dispersion, skewness, and kurtosis.

Creating a Pandas Series object:

CODE:

```
import pandas as pd
a=pd.Series([19,23,19,18,25,16,17,19,15,23,21,23,21,11,6])
```

The Pandas *describe* method can be used with either the series object or the DataFrame object and is a convenient way of obtaining most of the measures of central tendency in one line of code. The mean is 18.4 kgs, the first quartile (Q1 or 25th percentile) is 16.5 kgs, the median (50th percentile) is 19 kgs, and the third quartile (75th percentile) is 22 kgs.

CODE:

```
a.describe()
```

Output:

```
count    15.000000
mean     18.400000
std       4.997142
min       6.000000
25%      16.500000
50%      19.000000
75%      22.000000
max      25.000000
dtype: float64
```

Obtain the mode using the *mode* method:

CODE:

```
a.mode()
```

Output:

```
0    19
1    23
dtype: int64
```

The values 19 and 23 are the most frequently occurring values.

Obtain the measures of dispersion

The range can be calculated using the *max* and *min* functions and taking the difference between these two values:

CODE:

```
range_a=max(a)-min(a)
range_a
```

Output:

```
19
```

Obtain the standard deviation and variance using the *std* and *var* methods, respectively:

CODE:

```
a.std()
```

Output:

```
4.99714204034952
```

CODE:

```
a.var()
```

Output:

```
24.97142857142857
```

Obtain the measures of skewness and kurtosis by using the *skew* and *kurtosis* functions from the *scipy.stats* module:

CODE:

```
from scipy.stats import skew,kurtosis
stats.kurtosis(a)
```

Output:

```
0.6995494033062934
```

A positive value of kurtosis means that the distribution is leptokurtic.

Skewness:

CODE:

```
stats.skew(a)
```

Output:

```
-1.038344732097918
```

A negative value of skewness implies that the distribution is negatively skewed, with the mean less than the median.

Points to note:

1. The mean value is affected by outliers (extreme values). Whenever there are outliers in a dataset, it is better to use the median.
2. The standard deviation and variance are closely tied to the mean. Thus, if there are outliers, standard deviation and variance may not be representative measures too.
3. The mode is generally used for discrete data since there can be more than one modal value for continuous data.

Sampling

When we try to find out something about a population, it is not practical to collect data from every subject in the population. It is more feasible to take a sample and make an estimate of the population parameters based on the sample data. The sample's characteristics should align with that of the population. The following are the main methods of collecting samples.

Probability sampling

Probability sampling involves a random selection of subjects from the population. There are four main methods of doing probability sampling:

1. **Simple random sampling:** Subjects are chosen randomly, without any preference. Every subject in the population has an equal likelihood of being selected.
2. **Stratified random sampling:** The population is divided into mutually exclusive (non-overlapping) groups, and then subjects are randomly selected from each group. Example: If you are surveying to assess preference for subjects in a school, you may divide students into male and female groups and randomly select subjects from each group. The advantage of this method is that it represents all categories or groups in the population.

3. **Systematic random sampling:** Subjects are chosen at regular intervals. Example: To take a sample of 100 people from a population of 500, first divide 500 by 100, which equals 5. Now, take every 5th person for our sample. It is easier to perform but may not be representative of all the subjects in the population.
4. **Cluster sampling:** Here, the population is divided into non-overlapping clusters covering the entire population between them. From these clusters, a few are randomly selected. Either all the members of the chosen clusters are selected (one-stage), or a subset of members from the selected clusters is randomly chosen (two-stage). The advantage of this method is that it is cheaper and more convenient to carry out.

Non-probability sampling

When it is not possible to collect samples using probability sampling methods due to a lack of readily available data, we use non-probability sampling techniques. In non-probability sampling, we do not know the probability of a subject being chosen for the study.

It is divided into the following types:

1. **Convenience sampling:** Subjects that are easily accessible or available are chosen in this method. For example, a researcher can select subjects for their study from their workplace or the university where they work. This method is easy to implement but may not be representative of the population.
2. **Purposive:** Subjects are chosen based on the purpose of the sampling. For example, if a survey is being carried out to assess the effectiveness of intermittent fasting, it needs to consider the age group of the population that can undergo this fast, and the survey may only include people aged 25-50. Purposive sampling is further divided into
 - **Quota sampling:** Quotas are taken in such a way that the significant characteristics of the population are taken into account while samples are chosen. If a population has 60% Caucasians, 20% Hispanics, and 20% Asians, the sample you choose should have the same percentages.

- **Snowball sampling:** In this method, the researcher identifies someone they know who meets the criteria of the study. This person then introduces to others the person may know, and the sample group thus grows through word-of-mouth. This technique may be used for populations that lack visibility, for example, a survey of people suffering from an under-reported illness.

Central limit theorem

The central limit theorem states that if we choose samples from a population, the means of the samples are normally distributed, with a mean as μ and standard deviation as $\sigma_{\bar{x}}$.

Even if the population distribution is not a normal distribution by itself, the distribution of the sample means resembles a normal distribution. As the sample size increases, the distribution of sample means becomes a closer approximation to the normal distribution, as seen in Figure 9-9.

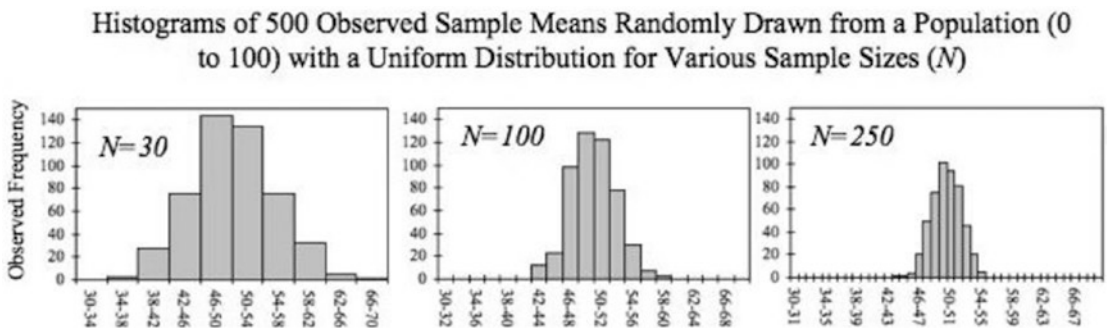


Figure 9-9. *Distribution of sample means. As the sample size increases, the distribution of sample means resembles a normal distribution*

The sample mean is used as an estimate for the population mean, but the standard deviation of this sampling distribution ($\sigma_{\bar{x}}$), is not the same as the population standard deviation, σ . The sample standard deviation is related to the population standard deviation as follows:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the population standard deviation and n is the sample size

$\sigma_{\bar{x}}$ is known as the standard error (of the distribution of sample means). As the sample size (n increases), the standard error approaches 0, and the sample mean (\bar{x}) approaches the population mean (μ).

Estimates and confidence intervals

Point estimate: A single statistic extracted from a sample that is used to estimate an unknown population parameter. The sample mean is used as a point estimate for the population mean.

Interval estimate: The broad range of values within which the population parameter lies. It is indicative of the error in estimating the population parameter.

Confidence interval: The interval within which the value of the population mean lies. For a random sample of size n and mean \bar{x} taken from a population (with standard deviation as σ , and mean as μ), the confidence interval for the population mean is given by the following equations:

$$\bar{x} - \frac{z\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{z\sigma}{\sqrt{n}} : \text{when population standard deviation, } \sigma, \text{ is known}$$

$$\bar{x} - \frac{zs}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{zs}{\sqrt{n}} : \text{when population standard deviation is unknown (s in this equation is the sample standard deviation)}$$

Solved example: Confidence intervals

Question: A sample (consisting of ten subjects) is taken from a certain population of students. The grade point averages of these students are normally distributed. The population standard deviation is not known. Calculate the 95% confidence interval for the population mean (grade point average for the whole student population), based on the following sample values: 3.1, 2.9, 3.2, 3.4, 3.7, 3.9, 3.9, 2.8, 3.4, 3.6.

Solution:

The following code calculates the 95% confidence interval for the population mean:

CODE:

```
import numpy as np
import scipy.stats as stats
grades = np.array([3.1,2.9,3.2,3.4,3.7,3.9,3.9,2.8,3.4,3.6])
stats.t.interval(0.95, len(grades)-1, loc=np.mean(grades), scale=stats.
sem(grades))
```

Output:

```
(3.1110006165952773, 3.668999383404722)
```

Interpretation: There is a 95% probability that the grade point average for the population of students falls between 3.11 and 3.67.

Explanation of the preceding code: We first define a NumPy array for the sample observations, and then call the *stats.t.interval* function. To this function, we pass the following arguments: the values of the confidence interval (0.95), degrees of freedom (total number of observations: 1), the sample mean, and standard error (calculated by the function *stats.sem*). The function returns two values – the lower confidence interval (LCI) and the upper confidence interval (UCI). Note that the *stats.t.interval* function is used because the population standard deviation is not known. If it were known, we would use the function *stats.norm.interval*.

Types of errors in sampling

If we take a sample and make inferences about the entire population based on this sample, errors inevitably arise. These errors can broadly be classified as follows:

- Sampling error: Difference between the sample estimate for the population and the actual population estimate
- Coverage error: Occurs when the population is not adequately represented, and some groups are excluded

- **Nonresponse errors:** Occurs when we fail to include nonresponsive subjects that satisfy the criteria of the study, but are excluded since they do not answer the survey questions
- **Measurement error:** Not measuring the correct parameters due to flaws in the method or tool used for measurement

We now move on to concepts in hypothesis testing.

Hypothesis testing

A hypothesis is a statement that gives the estimate of an unknown variable or parameter. If we are trying to find the average age of people in a city from a sample drawn from this population, and we find that the average age of people in this sample is 34, our hypothesis statement could be as follows: “The average age of people in this city is 34 years.”

Basic concepts in hypothesis testing

In a hypothesis test, we construct two statements known as the null and alternate hypothesis.

Null hypothesis: Denoted by the term H_0 , this is the hypothesis that needs to be tested. It is based on the principle that there is no change from the status quo. If the sample mean is 70, while the historical population mean is 90, the null hypothesis would state that the population mean equals 90.

Alternate hypothesis: Denoted by the term H_1 , this hypothesis is what one would believe if the null hypothesis does not hold. The alternate hypothesis (using the preceding example) would state that the mean is greater than, less than, or not equal to 90.

We either reject the null hypothesis or fail to reject the null hypothesis. Note that rejecting the null hypothesis does not imply that the alternative hypothesis is true. The result of a hypothesis test is only suggestive or indicative of something regarding the population, and it does not conclusively prove or disprove any hypothesis.

Key terminology used in hypothesis testing

Let us look at some commonly used terms in hypothesis testing:

Type 1 error, or the **level of significance**, denoted by the symbol α , is the error of rejecting the null hypothesis when it is true. It can also be defined as the probability that the population parameter lies outside its confidence interval. If the confidence interval is 95%, the level of significance is 0.05, or there is a 5% chance that the population parameter does not lie within the confidence interval calculated from the sample.

Example of a Type 1 error: Mr. X has a rash and goes to a doctor to get a test for chickenpox. Let the null hypothesis be that he does not have this illness. The doctor incorrectly makes a diagnosis for chickenpox based on some faulty tests, but the reality is that Mr. X does not have this illness. This is a typical example of rejecting the null hypothesis when it is true, which is what a Type 1 error is.

Type 2 error, denoted by the symbol β , is the error that occurs when the null hypothesis is not rejected when it is false. In the preceding chickenpox example, if Mr. X suffers from chickenpox, but the doctor does not diagnose it, the doctor is making a Type 2 error.

One-sample test: This is a test used when there is only one population under consideration, and a single sample is taken to see if there is a difference between the values calculated from the sample and population parameter.

Two-sample test: This is a test used when samples are taken from two different populations. It helps to assess whether the population parameters are different based on the sample parameters.

The critical test statistic: The limiting value of the sample test statistic to decide whether or not to reject the null hypothesis. In Figure 9-10, $z=1.96$ and $z=-1.96$ are critical values. Z-values greater than 1.96 and less than -1.96 lead to rejection of the null hypothesis.

Region of rejection: The range of values where the null hypothesis is rejected. The region of acceptance is the area corresponding to the limits where the null hypothesis holds.

The regions of rejection and acceptance are shown in Figure 9-10.

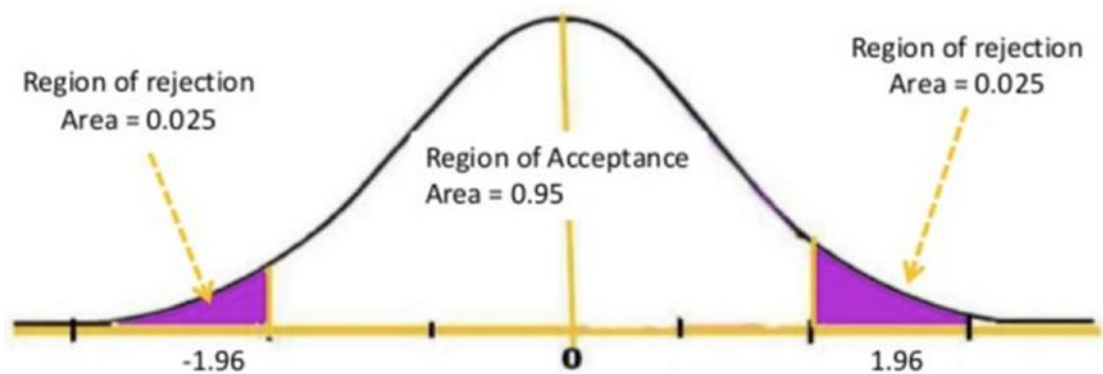


Figure 9-10. Regions of acceptance and rejection

Two-tail test: The region of rejection is located on both tails of the distribution.

Example of a two-tail test: A sample of 10 students is taken from a class of 50 students to see if there is a change in the mean score of the class with respect to its historical average. This is an example of a case where we will conduct a two-tail test because we are just testing for a change in the mean, and we do not know if this change is positive or negative.

One-tail test: The region of rejection is located on the right tail (upper one-tail) or the left tail (lower-tail) but not on both tails.

Example of upper-tail test: Afterschool classes are being conducted to improve scores in a class of 50 students. These special classes are believed to have improved the scores. To test this hypothesis, we perform a one-tail test (upper) using a sample from the population because we are testing if the mean score has increased. The region of rejection will be located on the right tail.

Example of lower-tail test: Due to political unrest, there has been an increase in absenteeism among students. It is believed that these events may negatively affect the scores of the students. To test this hypothesis, we conduct a one-tail test (lower) using a sample from the population because we are testing if the mean score has reduced. The region of rejection is located on the left tail.

The p-value (denoted by the letter p) is the probability of obtaining a value of the test statistic at least as extreme as the one observed, assuming the null hypothesis is true.

The p-value is often used in hypothesis tests to decide whether or not to reject a null hypothesis. The p-value is commonly compared with a significance level of 0.05.

If $p < 0.05$, it would mean that the probability that the sample data was random and not representative of the population is very low. We reject the null hypothesis in this case.

If $p > 0.05$, there is a greater chance that this sample is not representative of the population. We fail to reject the null hypothesis in this case.

Steps involved in hypothesis testing

1. State the null and alternate hypothesis
2. Fix the level of significance and obtain the critical value of the test statistic
3. Select the appropriate test:

Choose the test based on the following parameters:

- Number of samples
 - Whether the population is normally distributed
 - The statistic being tested
 - The sample size
 - Whether the population standard deviation is known
4. Obtain the relevant test statistic (z statistic/t statistic/chi-square statistic/f statistic) or the p-value
 5. Compare the critical test statistic with the calculated test static or compare the p-value with 0.05

Reject the null hypothesis based on either the test statistic or the p-value:

- Using the test statistic:
 - calculated test static > critical test statistic (upper-tail test)
 - calculated test static < critical test statistic (lower-tail test)

OR

- Using the p-value (p) if $p < 0.05$
6. Draw an inference based on the preceding comparison

One-sample z-test

This test is used when we want to verify if the population mean differs from its historical or hypothesized value.

Criteria for a one-sample z-test:

- The population from which the sample is drawn is normally distributed
- The sample size is greater than 30
- A single sample is drawn
- We are testing for the population mean
- The population standard deviation is known

Formula for calculating test statistic: $z = \frac{(\bar{x} - \mu)}{\sigma / \sqrt{n}}$,

where \bar{x} is the sample mean, μ is the population mean, σ is the population standard deviation, and n is the sample size

Solved example: One-sample z-test

Question: A local Italian restaurant has an average delivery time of 45 minutes with a standard deviation of 5 minutes. The restaurant has received some complaints from its customers and has decided to analyze the last 40 orders. The average delivery time for these 40 orders was found to be 48 minutes. Conduct the appropriate test at a significance level of 5% to decide whether the delivery times have increased.

Answer:

1. State the hypothesis:

Let μ be the average delivery time for the restaurant
(population mean)

Null hypothesis: $H_0: \mu=45$

Alternate hypothesis: $H_1: \mu > 45$

2. Fix the level of significance: $\alpha=0.05$

3. Select the appropriate hypothesis test:

- Number of samples: 1
- Sample size: $n=40$ (Large)
- What we are testing: Whether there is a difference between the sample mean ($\bar{x} = 48$) and the population mean ($\mu=45$)
- Population standard deviation ($\sigma=5$) is known

We select the one-sample z-test based on the preceding data.

4. Obtain the test statistic and p-value, with the help of the following equation:

$$z = \frac{(\bar{x} - \mu)}{\sigma / \sqrt{n}}$$

Substituting the values $\bar{x} = 48$, $\mu=45$, $\sigma = 5$, and $n=40$:

$$z=3.7947$$

Calculate the p-value corresponding to this z-value using the *stats.norm.cdf* function:

CODE:

```
import scipy.stats as stats
stats.norm.cdf(z)
```

Output:

0.999

5. Compare the p-value with the level of significance (0.05):

Since the calculated p-value is $> \alpha$, we fail to reject the null hypothesis.

6. Inference:

There is no significant difference, at a level of 0.05, between the average delivery time of the sample and the historical population average.

Two-sample sample z-test

A two-sample z-test is similar to a one-sample z-test, the only differences being as follows:

- There are two groups/populations under consideration and we draw one sample from each population
- Both the population distributions are normal
- Both population standard deviations are known
- The formula for calculating test statistic: $z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right]}}$

Solved example: Two-sample sample z-test

An organization manufactures LED bulbs in two production units, A and B. The quality control team believes that the quality of production at unit A is better than that of B. Quality is measured by how long a bulb works. The team takes samples from both units to test this. The mean life of LED bulbs at units A and B are 1001.3 and 810.47, respectively. The sample sizes are 40 and 44. The population variances are known: $\sigma_A^2 = 48127$ and $\sigma_B^2 = 59173$.

Conduct the appropriate test, at 5% significance level, to verify the claim of the quality control team.

Solution:

1. State the hypothesis:

Let the mean life of LED bulbs at unit A and B be μ_A and μ_B , respectively.

Null hypothesis: $H_0: \mu_A \leq \mu_B$

Alternate hypothesis: $H_1: \mu_A > \mu_B$

This is a one-tail (upper-tail) test

2. Fix the level of significance: $\alpha=0.05$

3. Select the appropriate hypothesis test:

- Number of samples: 2 samples (taking samples from two different populations)
- Sample size: Large ($n_A = 40$, and $n_B = 44$)
- What we are testing: Comparing the mean lifetime of LED bulbs in unit A with that of unit B
- Population characteristics: The distribution of population is not known, but population variances are known
- Hence, we conduct the two-sample z-test.

4. Calculate the test statistic and p-value

Use the following equation:

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right]}}$$

Substituting the values $\bar{x}_1 = 1001.3, \bar{x}_2 = 810.47, n_1 = 40, n_2 = 44$ and the variance(sigma) values of 48127 and 59173 in the preceding formula to calculate z:

CODE:

```
z=(1001.34-810.47)/(48127/40+59173/44)**0.5
```

Output:

```
3.781260568723408
```

Calculate the p-value corresponding to this z-value using the *stats.norm.cdf* function:

CODE:

```
import scipy.stats as stats
p=1-stats.norm.cdf(z)
p
```

Output:

```
7.801812433294586e-05
```

Explanation: Since this is an upper-tail test, we need to calculate the area/proportion of values in the right tail. Hence, we subtract the area calculated (*stats.norm.cdf*) from 1.

5. Comparing the calculated p-value with the level of significance:

Since the calculated p-value ($0.000078 < \alpha(0.05)$), we reject the null hypothesis.

6. Inference: The LED bulbs produced at unit A have a significantly longer life than those at unit B, at a 5% level.

Hypothesis tests with proportions

Proportion tests are used with nominal data and are useful for comparing percentages or proportions. For example, a survey collecting responses from a department in an organization might claim that 85% of people in the organization are satisfied with its policies. Historically the satisfaction rate has been 82%. Here, we are comparing a percentage or a proportion taken from the sample with a percentage/proportion from the population. The following are some of the characteristics of the sampling distribution of proportions:

- The sampling distribution of the proportions taken from the sample is approximately normal
- The mean of this sampling distribution (\bar{p}) = Population proportion (p)
- Calculating the test statistic: The following equation gives the z-value

$$z = \frac{(\bar{p} - p)}{\sqrt{\frac{p(1-p)}{n}}}$$

Where \bar{p} is the sample proportion, p is the population proportion, and n is the sample size.

Solved example: One-sample proportion z-test

Here, we understand the one-sample proportion z-test using a solved example.

Question: It is known that 40% of the total customers are satisfied with the services provided by a mobile service center. The customer service department of this center decides to conduct a survey for assessing the current customer satisfaction rate. It surveys 100 of its customers and finds that only 30 out of the 100 customers are satisfied with its services. Conduct a hypothesis test at a 5% significance level to determine if the percentage of satisfied customers has reduced from the initial satisfaction level (40%).

Solution:

1. State the null and alternate hypothesis

Let the average customer satisfaction rate be p

$$H_0: p = 0.4$$

$$H_1: p < 0.4$$

The $<$ sign indicates that this is a one-tail test (lower-tail)

2. Fix the level of significance: $\alpha=0.05$
3. Select the appropriate test:

We choose the one-sample z-test for proportions since

- The sample size is large (100)
- A single sample is taken
- We are testing for a change in the population proportion

4. Obtain the relevant test statistic and p-value

$$z = \frac{(\bar{p} - p)}{\sqrt{\frac{p(1-p)}{n}}}$$

Where $\bar{p} = 0.3, p = 0.4, n = 100$

Calculate z and p:

CODE:

```
import scipy.stats as stats
z=(0.3-0.4)/((0.4)*(1-0.4)/100)**0.5
p=stats.norm.cdf(z)
p
```

Output:

0.02061341666858179

5. Decide whether or not to reject the null hypothesis
p-value (0.02)<0.05 → We reject the null hypothesis
6. Inference: At a 5% significance level, the percentage of customers satisfied with the service center's services has reduced

Two-sample z-test for the population proportions

Here, we compare proportions taken from two independent samples belonging to two different populations. The following equation gives the formula for the critical test statistic:

$$z = \frac{(\bar{p}_1 - \bar{p}_2)}{\sqrt{\frac{p_c(1-p_c)}{N_1} + \frac{p_c(1-p_c)}{N_2}}}$$

In the preceding formula, \bar{p}_1 is the proportion from the first sample, and \bar{p}_2 is the proportion from the second sample. N_1 is the sample size of the first sample, and N_2 is the sample size of the second sample.

p_c is the pooled variance.

$$\bar{p}_1 = \frac{x_1}{N_1}; \bar{p}_2 = \frac{x_2}{N_2}; p_c = \frac{x_1 + x_2}{N_1 + N_2}$$

In the preceding formula, x_1 is the number of successes in the first sample, and x_2 is the number of successes in the second sample.

Let us understand the two-sample proportion test with the help of an example.

Question: A ride-sharing company is investigating complaints by its drivers that some of the passengers (traveling with children) do not conform with child safety guidelines (for example, not bringing a child seat or not using the seat belt). The company undertakes surveys in two major cities. The surveys are collected independently, with one sample being taken from each city. From the data collected, it seems that the passengers in City B are more noncompliant than those in City A. The law enforcement authority wants to know if the proportion of passengers conforming with child safety guidelines is different for the two cities. The data for the two cities is given in the following table:

	City A	City B
Total surveyed	200	230
Number of people compliant	110	106

Conduct the appropriate test, at 5% significance level, to test the hypothesis.

Solution:

1. State the hypothesis:

Let p_A be the proportion of people in City A who are compliant with the norms and p_B be the proportion of people in City B who are compliant with the standards.

Null hypothesis: $H_0: p_A = p_B$

Alternate hypothesis: $H_1: p_A \neq p_B$

This would be a two-tail test, because the region of rejection could be located on either side.

2. Select the appropriate hypothesis test:
 - Number of samples: 2 (taking samples from two different cities)
 - Sample size: Large ($N_1 = 200$ and $N_2 = 230$)
 - What we are testing: Whether the proportion of passengers conforming with child safety guidelines is different for the two cities
 - Population characteristics: The distribution of the population is not known; population variances are unknown. Since sample sizes are large, we select the two-sample z-test for proportions

3. Fix the level of significance: $\alpha=0.05$
4. Calculate the test statistic and p-value

Using the following equation:

$$z = \frac{(\bar{p}_1 - \bar{p}_2)}{\sqrt{\frac{p_c(1-p_c)}{N_1} + \frac{p_c(1-p_c)}{N_2}}}$$

Calculate the p-value corresponding to this z-value using the *stats.norm.cdf* function:

CODE:

```
x1,n1,x2,n2=110,200,106,230
p1=x1/n1
p2=x2/n2
pc=(x1+x2)/(n1+n2)
z=(p1-p2)/(((pc*(1-pc)/n1)+(pc*(1-pc)/n2))**.5)
p=2*(1-stats.norm.cdf(z))
p
```

Output:

0.06521749465064053

5. Comparing the p-value with the level of significance:

Since the calculated p-value $(0.065) > \alpha(0.05)$, we fail to reject the null hypothesis.

6. Inference: There is no significant difference between the proportion of passengers in these cities complying with child safety norms, at a 5% significance level.

T-distribution

There may be situations where the standard deviation of the population is unknown, and the sample size is small. In such cases, we use the T-distribution. This distribution is also called Student's T distribution. The word "Student" does not assume its literal

meaning here. William Sealy Gosset, who first published this distribution in 1908, used the pen name “Student,” and thus this distribution became widely known as Student’s T-distribution.

The following are the chief characteristics of the T-distribution:

- The T-distribution is similar in shape to a normal distribution, except that it is slightly flatter.
- The sample size is small, generally less than 30.
- The T-distribution uses the concept of degrees of freedom. The degrees of freedom are the number of observations in a statistical test that can be estimated independently. Let us understand the concept of degrees of freedom using the following example:

Say we have three numbers: a, b, and c. We do not know their values, but we know the mean of the three numbers, which is 5. From this mean value, we calculate the sum of the three numbers – 15 (mean*number of values, 5*3).

Can we assign any value to these three unknown numbers? No; only two of these three numbers can be assigned independently. Say we randomly assign the value 4 to a and 5 to b. Now, c can only be 6 since the total sum has to be 15. Hence, even though we have three numbers, only two are free to vary.

- As the sample size decreases, the degrees of freedom reduce, or in other words, the certainty with which the population parameter can be predicted from the sample parameter reduces.

The degrees of freedom (df) in the T-distribution is the number of samples (n) -1, or in other words, $df = n - 1$.

The formula for the critical test statistic in a one-sample t-test is given by the following equation:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

where \bar{x} is the sample mean, μ is the population mean, s is the sample standard deviation and n is the sample size.

One sample t-test

A one-sample t-test is similar to a one-sample z-test, with the following differences:

1. The size of the sample is small (<30).
2. The population standard deviation is not known; we use the sample standard deviation(s) to calculate the standard error.
3. The critical statistic here is the t-statistic, given by the following formula:

$$t = \frac{(\bar{x} - \mu)}{s / \sqrt{n}}$$

Two-sample t-test

A two-sample t-test is used when we take samples from two populations, where both the sample sizes are less than 30, and both the population standard deviations are unknown.

Formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Where \bar{x}_1 and \bar{x}_2 are the sample means

The degrees of freedom: $df = n_1 + n_2 - 2$

The pooled variance: $S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$

Two-sample t-test for paired samples

This test is used to compare population means from samples that are dependent on each other, that is, sample values are measured twice using the same test group.

This equation gives the critical value of the test statistic for a paired two-sample t-test:

$$t = \frac{\bar{d}}{s / \sqrt{n}}$$

Where \bar{d} is the average of the difference between the elements of the two samples. Both the samples have the same size, n .

S = standard deviation of the differences between the elements of the two samples =

$$\sqrt{\frac{\sum d^2 - (\sum d)^2 / n}{n - 1}}$$

Solved examples: Conducting t-tests using Scipy functions

The Scipy library has various functions for the t-test. In the following examples, we look at the functions for the one-sample t-test, the two-sample t-test, and the paired t-test.

1. One-sample t-test with Scipy:

Question: A coaching institute, preparing students for an exam, has 200 students, and the average score of the students in the practice tests is 80. It takes a sample of nine students and records their scores; it seems that the average score has now increased. These are the scores of these ten students: 80, 87, 80, 75, 79, 78, 89, 84, 88.

Conduct a hypothesis test at a 5% significance level to verify if there is a significant increase in the average score.

Solution:

We use the one-sample t-test since the sample size is small, and the population standard deviation is not known. Let us formulate the null and alternate hypotheses.

$$H_0: \mu = 80$$

$$H_1: \mu > 80$$

First, create a NumPy array with the sample observations :

CODE:

```
a=np.array([80,87,80,75,79,78,89,84,88])
```

Now, call the *stats.ttest_1samp* function and pass this array and the population mean. This function returns the t-statistic and the p-value.

CODE:

```
stats.ttest_1samp(a,80)
```

Output:

```
Ttest_1sampResult(statistic=1.348399724926488, pvalue=0.21445866072113726)
```

Decision: Since the p-value is greater than 0.05, we fail to reject the null hypothesis. Hence, we cannot conclude that the average score of students has changed.

2. Two-sample t-test (independent samples):

Question: A coaching institute has centers in two different cities. It takes a sample of ten students from each center and records their scores, which are as follows:

Center A: 80, 87, 80, 75, 79, 78, 89, 84, 88

Center B: 81, 74, 70, 73, 76, 73, 81, 82, 84

Conduct a hypothesis test at a 5% significance level, and verify if there is a significant difference in the average scores of the students in these two centers.

Solution:

We use the two-sample t-test since we are taking samples from two independent groups. The sample size is small, and the standard deviations of the populations are not known. Let the average scores of students in each of these centers be μ_1 and μ_2 . The null and alternate hypothesis is as follows:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Create NumPy arrays for each of these samples:

CODE:

```
a=np.array([80,87,80,75,79,78,89,84,88])
b=np.array([81,74,70,73,76,73,81,82,84])
```

Call the *stats.ttest_ind* function to conduct the two-sample t-test and pass these arrays as arguments:

CODE:

```
stats.ttest_ind(a,b)
```

Output:

```
Ttest_indResult(statistic=2.1892354788555664,
pvalue=0.04374951024120649)
```

Inference: We can conclude that there is a significant difference in the average scores of students in the two centers of the coaching institute since the p-value is less than 0.05.

3. T-test for paired samples:

Question: The coaching institute is conducting a special program to improve the performance of the students. The scores of the same set of students are compared before and after the special program. Conduct a hypothesis test at a 5% significance level to verify if the scores have improved because of this program.

Solution:

CODE:

```
a=np.array([80,87,80,75,79,78,89,84,88])
```

```
b=np.array([81,89,83,81,79,82,90,82,90])
```

Call the *stats.ttest_rel* function to conduct the two-sample t-test and pass these arrays as arguments:

CODE:

```
stats.ttest_rel(a,b)
```

Output:

```
Ttest_relResult(statistic=-2.4473735525455615,  
pvalue=0.040100656419513776)
```

We can conclude, at a 5% significance level, that the average score has improved after the special program was conducted since the p-value is less than 0.05.

ANOVA

ANOVA is a method used to compare the means of more than two populations. So far, we have considered only a single population or at the most two populations. The statistical distribution used in ANOVA is the F distribution, whose characteristics are as follows:

1. The F-distribution has a single tail (toward the right) and contains only positive values, as shown in Figure 9-11.

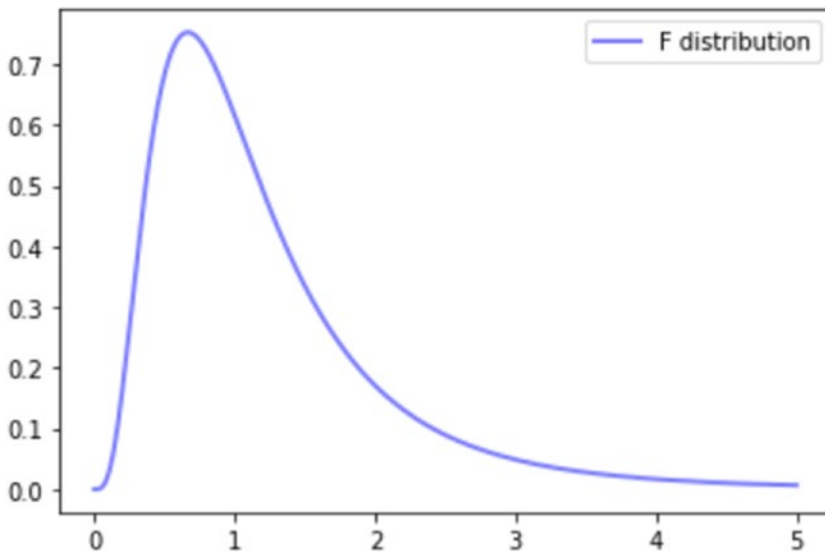


Figure 9-11. Shape of the F-distribution

2. The F-statistic, which is the critical statistic in ANOVA, is the ratio of variation between the sample means to the variation within the sample. The formula is as follows.

$$F = \frac{\text{variation between sample means}}{(\text{variation within the samples})}$$

3. The different populations are referred to as treatments.
4. A high value of the F statistic implies that the variation between samples is considerable compared to variation within the samples. In other words, the populations or treatments from which the samples are drawn are actually different from one another.
5. Random variations between treatments are more likely to occur when the variation within the sample is considerable.

Solved example: ANOVA

Question:

A few agricultural research scientists have planted a new variety of cotton called “AB cotton.” They have used three different fertilizers – A, B, and C – for three separate plots of this variety. The researchers want to find out if the yield varies with the type of fertilizer used. Yields in bushels per acre are mentioned in the below table. Conduct an ANOVA test at a 5% level of significance to see if the researchers can conclude that there is a difference in yields.

Fertilizer A	Fertilizer B	Fertilizer C
40	45	55
30	35	40
35	55	30
45	25	20

Solution:

- 1. State the null and alternative hypothesis:

Let the average yields of the three populations be μ_1, μ_2 and μ_3

Null hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3$

Alternative hypothesis: $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$

- 2. Select the appropriate test:

We select the ANOVA test because we are comparing averages from three populations

- 3. Fix the level of significance: $\alpha=0.05$

- 4. Calculate the critical test statistic/p-value:

The `f_oneway` function gives us the test statistic or the p-value for the ANOVA test. The arguments to this function include three lists containing sample values of each of the groups.

CODE:

```
import scipy.stats as stats

a=[40,30,35,45]
b=[45,35,55,25]
c=[55,40,30,20]

stats.f_oneway(a,b,c)
```

Output:

```
F_onewayResult(statistic=0.10144927536231883,
pvalue=0.9045455407589628)
```

5. Since the calculated p-value (0.904)>0.05, we fail to reject the null hypothesis.
6. Inference: There is no significant difference between the three treatments, at a 5% significance level.

Chi-square test of association

The chi-square test is a nonparametric test for testing the association between two variables. A non-parametric test is one that does not make any assumption about the distribution of the population from which the sample is drawn. Parametric tests (which include z-tests, t-tests, ANOVA) make assumptions about the distribution/shape of the population from which the sample is drawn, assuming that the population is normally distributed. The following are some of the characteristics of the chi-square test.

- The chi-square test of association is used to test if the frequency of occurrence of one categorical variable is significantly associated with that of another categorical variable.
- The chi-square test statistic is given by: $X^2 = \frac{\sum (f_o - f_e)^2}{f_e}$, where f_o denotes the observed frequencies, f_e denotes the expected frequencies, and X is the test statistic. Using the chi-square test of association, we can assess if the differences between the frequencies are statistically significant.

- A contingency table is a table with frequencies of the variable listed under separate columns. The formula for the degrees of freedom in the chi-square test is given by:

$df=(r-1)*(c-1)$, where df is the number of degrees of freedom, r is the number of rows in the contingency table, and c is the number of columns in the contingency table.

- The chi-square test compares the observed values of a set of variables with their expected values. It determines if the differences between the observed values and expected values are due to random chance (like a sampling error), or if these differences are statistically significant. If there are only small differences between the observed and expected values, it may be due to an error in sampling. If there are substantial differences between the two, it may indicate an association between the variables.
- The shape of the chi-square distribution for different values of k (degrees of freedom) is shown in Figure 9-12. The chi-square distribution's shape varies with the degrees of freedom (denoted by k in Figure 9-12). When the degrees of freedom are few, it looks like an F-distribution. It has only one tail (toward the right). As the degrees of freedom increase, it looks like a normal curve. Also, the increase in the degrees of freedom indicates that the difference between the observed values and expected values could be meaningful and not just due to a sampling error.

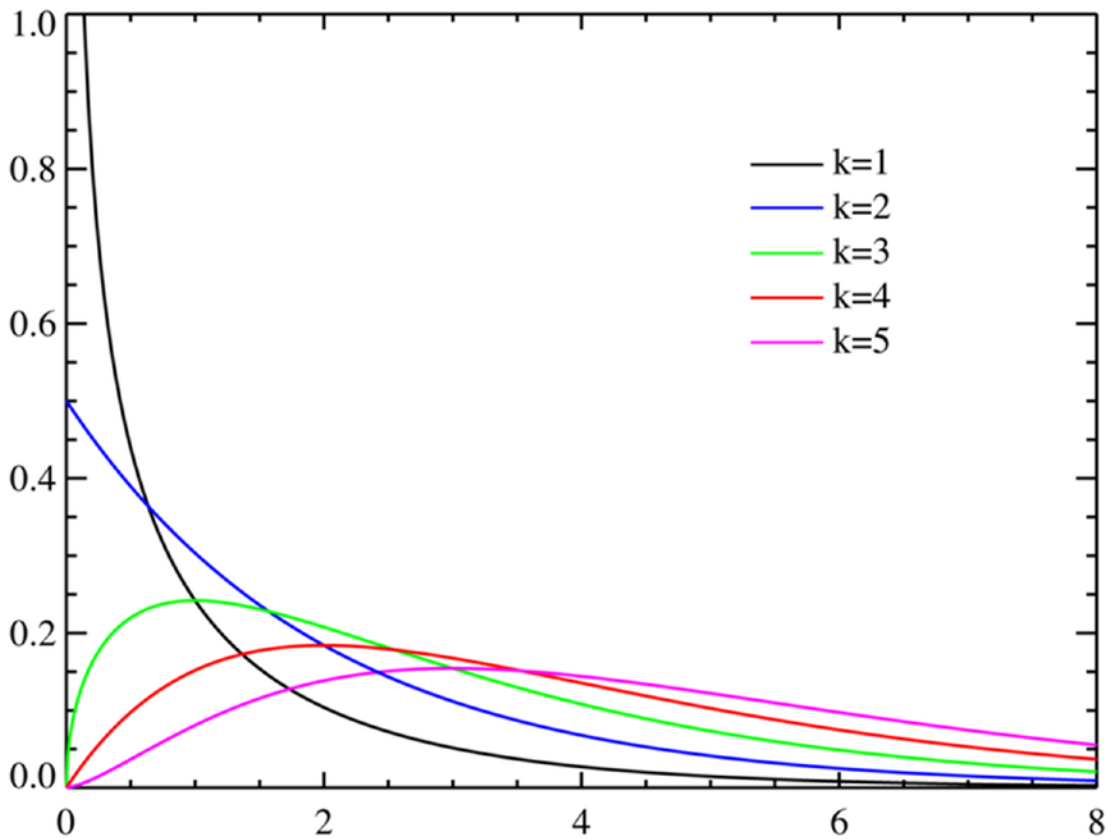


Figure 9-12. *Chi-square distribution for different degrees of freedom*

Solved example: Chi-square test

Question: A career counseling service guides students to help them understand their strengths and weaknesses so that they make appropriate career choices. They would like to assess if there is an association between the gender of a student and the career he or she chooses. The following table shows the number of males and females, and the careers (given by career IDs like I001, I002, etc.) they choose to pursue.

Career	Males	Females	Total
I001	41	79	120
I002	32	28	60
I003	58	78	130
I004	59	31	90

Answer:

1. State the hypothesis:
 - Null hypothesis: H_0 : gender and career preference are not related
 - Alternative hypothesis: H_1 : gender and career preference are related
2. Select the appropriate hypothesis test:
 - Number of variables: two categorical variables (gender and career)
 - What we are testing: Testing for an association between career and gender

We conduct a chi-square test of association based on the preceding characteristics.

3. Fix the level of significance: $\alpha=0.05$
4. Calculate the test statistic and p-value. The *chi2_contingency* function calculates the test statistic and p-value. This function returns the test statistic, the p-value, the degrees of freedom, and the expected frequencies (in the form of an array). The arguments to this function are the observations from the contingency table in the form of arrays. Each array represents a row in the contingency table.

CODE:

```
import scipy.stats as stats
observations=np.array([[41,79],[32,28],[52,78],[59,31]])
chi2stat,pval,dof,expvalue=stats.chi2_contingency(observations)
print(chi2stat,pval,dof,expvalue)
```

Output:

```
23.803453211665776 2.7454871071500803e-05 3 [[55.2 64.8]
 [27.6 32.4]
 [59.8 70.2]
 [41.4 48.6]]
```

The highlighted value in the preceding output is the p-value for this test.

5. Comparing the p-value with the level of significance:

Since the calculated p-value ($0.000027 < \alpha(0.05)$), we reject the null hypothesis.

6. Inference: There is a significant association between the gender of the student and career choice, at a 5% significance level.

Caveat while using p-value:

The power of a hypothesis test is measured by its ability to yield statistically significant results, which is represented by a p-value that is less than 0.05. The results of many research trials and experiments conducted in the fields of medical and social sciences are presented using p-values. The p-value, however, is hard to interpret. It is also dependent on the sample size and the size of the bias that we measure. A result that is statistically significant does not conclusively disprove the null hypothesis or prove the alternate hypothesis. Confidence intervals are generally preferable to using p-values, as they are more easily interpretable.

Summary

1. Combinations refer to the number of ways in which we can select items, whereas permutations refer to the number of ways in which we can arrange them.

2. Probability is the likelihood of an event.

Two events are independent when the probability of occurrence of one event does not affect the other. Independent events follow the special rule of multiplication, where $P(A \cap B) = P(A) * P(B)$.

Mutually exclusive events are those that cannot occur together, and such events follow the special rule of addition, where $P(A \cup B) = P(A) + P(B)$.

3. The Bayes theorem calculates the posterior probability of an event, or in other words, the probability of a hypothesis being true given some evidence related to it.

4. A random variable takes the values associated with the outcomes of an experiment. There are two types of random variables: discrete (which take only a few values) and continuous (which can take any number of values).
5. Discrete variables can be used for binomial distributions (a single experiment repeated multiple times with each trial having two possible outcomes) or Poisson distributions (which model the number of occurrences that occur over an interval, given the average rate of occurrence).
6. The normal distribution is a symmetric bell-shaped curve, using a continuous random variable with most of its values centered around the mean. The standard normal distribution has a mean of 0 and a standard deviation of 1. The formula used in standard normal distributions is as follows:

$$z = \frac{(x - \mu)}{\sigma} .$$

7. Continuous distributions have various measures of central tendency (mean, median, mode), dispersion (range, variance, standard deviation), and shape (skewness is a measure of asymmetry while kurtosis is a measure of the curvedness of the distribution).
8. A sample is used when it is impractical to collect data about all the subjects in a large population. The main methods of collecting a sample are probability sampling (where subjects are randomly selected from a large population) and non-probability sampling (when data is not readily available, and samples are taken based on availability or access).
9. A hypothesis test is used to make an inference about a population based on a sample, but it does not conclusively establish anything about the population. It is only suggestive. The two kinds of estimates that can be made about a population from a sample are point estimates (using a single value) and interval estimates (using

a range of values). The confidence interval, which is the range of values within which the population mean lies, is an example of an interval estimate.

10. The null hypothesis indicates that nothing has changed, while the alternate hypothesis is used when we have reason to reject the null hypothesis. A Type 1 error occurs when the null hypothesis is incorrectly rejected when it is true. In contrast, a Type 2 error occurs when we fail to reject the null hypothesis when it is not true.
11. Either the test statistic (which is different for every hypothesis test) or the p-value can be used to decide whether or not to reject the null hypothesis. The p-value measures the likelihood that the observed data occurred merely by chance.
12. A two-tail test is used when we are testing whether the population parameter(s) is not equal to a particular value. In contrast, a one-tail test is used when the population parameter(s) is either greater than or less than a particular value.

A one-sample test is used when a single sample is taken from a population, while a two-sample test compares samples taken from different populations.

13. Hypothesis tests can be either parametric (when we assume that the population from which a sample is drawn is normally distributed) or nonparametric (when we do not make such assumptions about the population distribution).
14. A parametric hypothesis test could be used to compare means using the z-test (when the sample size is large, and the population standard deviation is known) or the t-test (small sample size <30 and the population standard deviation is unknown). Z-tests can also be used to compare proportions. The ANOVA test is used when we need to compare the means of more than two populations. The chi-square test, one commonly used nonparametric test, is used for testing the association between variables.

Review Exercises

Question 1

Match the Scipy function (column on the right) with the appropriate hypothesis test (column on the left).

Hypothesis test	Scipy function
1. Chi-square	a. stats.ttest_rel
2. ANOVA	b. stats.ttest_1samp
3. Paired t-test	c. stats.f_oneway
4. One-sample t-test	d. stats.chi2_contingency
5. Two-sample (independent) t-test	e. stats.ttest_ind

Question 2

Skewness is a measure of:

- 1. Dispersion
- 2. Central tendency
- 3. Curvedness
- 4. Asymmetry

Question 3

Mr. J underwent a test for a widespread pandemic. The doctor made a clinical diagnosis that Mr. J does not have this illness. Later, when a blood test was conducted, it came out positive. Which of the following errors has the doctor committed?

- 1. Type 0 error
- 2. Type 1 error
- 3. Type 2 error
- 4. No error was committed

Question 4

Which of the following is correct?

1. A normal curve is an example of a mesokurtic curve
2. A platykurtic curve is flat
3. A leptokurtic curve has a high peak
4. All of the above
5. None of the above

Question 5

Let us assume that you are testing the effectiveness of e-learning programs in improving the score of students. The average score of the students is measured before and after the introduction of the e-learning programs. After comparing the means using a hypothesis test, you obtain a p-value of 0.02. This means that

1. The probability of the null hypothesis being true is 2%.
2. You have definitively disproved the null hypothesis (which states that there is no difference between the average scores before and after the introduction of the e-learning programs).
3. There is a 2% probability of getting a result as extreme as or more extreme than what has been observed.
4. You have definitively proved the alternative hypothesis.

Question 6

A new health drink claims to have 100 calories. The company manufacturing conducts periodic quality control check by selecting random independent samples (100 calories). The most recent 13 samples of this drink show the following calorie values: 78, 110, 105, 72, 88, 107, 85, 92, 82, 92, 91, 82, 103. At a significance level of 5%, conduct a hypothesis test whether there is a change in the calorific value of the health drink from what was originally claimed.

Question 7

Silver Gym is offering a fitness-cum-weight loss program for its clients and claims that this program will result in a minimum weight loss of 3 kgs after 30 days. To verify this claim, 20 clients who joined this program were studied. Their weights were compared before and after they underwent this program.

The weights of the 20 clients before and after the fitness program are as follows:

before_weights=[56,95,78,67,59,81,60,56,70,78,84,71,90,101,54,60]

after_weights=[52,91,77,65,54,78,54,55,65,76,82,66,88,94,53,55]

Conduct the appropriate test to test the hypothesis that there is a 3-kg weight loss (assuming that the weights of the population are normally distributed).

Answers

Question 1

1-d; 2-c; 3-a; 4-b; 5-e

Question 2

Option 4: Asymmetry (skewness is a measure of asymmetry)

Question 3

Option 3: Type 2 error

A Type 2 error is committed when the null hypothesis is not rejected when it does not hold true. Here, the null hypothesis is that the patient does not have this illness. The doctor should have rejected the null hypothesis and made a diagnosis for this illness since the blood test result is positive.

Question 4

Option 4: All of the above

Question 5

Option 3

Remember that the p-value only gives us the probability of getting a result as extreme as or more extreme than what is observed. It does not prove or disprove any hypothesis.

Question 6

1. State the hypothesis:

Let the mean calorie value of this drink be μ

Null hypothesis: $H_0: \mu=100$

Alternative hypothesis: $H_1: \mu \neq 100$

This is a two-tail test.

2. Select the appropriate hypothesis test:

We select the one-sample t-test based on the following characteristics:

- Number of samples: one sample
- Sample size: small ($n=13$)
- What we are testing: mean calorific value
- Population characteristics: population is normally distributed, and the population standard deviation is not known

3. Fix the level of significance: $\alpha=0.05$

4. Calculate the test statistic and p-value:

CODE:

```
import numpy as np
import scipy.stats as stats
values=np.array([78,110,105,72,88,107,85,92,82,92,91,82,103])
stats.ttest_1samp(values,100)
```

Output:

```
Ttest_1sampResult(statistic=-2.6371941582527527,
pvalue=0.02168579243588164)
```

5. Comparison: Since the calculated p-value $< \alpha$, we reject the null hypothesis.
6. Inference: It can be concluded, at a 5% level, that there is a significant difference between the calorific value of the sample and that of the population.

Question 7

1. State the hypothesis:

Let μ_d be the average difference in weights before and after the weight loss program for the population

Null hypothesis: $H_0: \mu_d < 3$

Alternative hypothesis: $\mu_d \geq 3$

One-tail test since there is a greater-than-or-equal-to sign in the alternative hypothesis

2. Select the appropriate hypothesis test:

- Number of samples: Two samples (taking two different samples with the same subjects)
- Sample size: Small (20)
- What we are testing: Testing the average difference in weight loss
- Population characteristics: Distribution of population is normal, but population variances are not known.

Based on the preceding characteristics and since the samples are related to each other (considering that we are comparing the weights of the same clients), we conduct a paired two-sample t-test.

3. Fix the level of significance: $\alpha=0.05$
4. Calculate the p-value:

The p-value can be calculated using the `stats.ttest_rel` equation as shown in the following code.

CODE:

```
import scipy.stats as stats
before_weights=[56,95,78,67,59,81,60,56,70,78,84,71,90,101,54,60]
after_weights=[52,91,77,65,54,78,54,55,65,76,82,66,88,94,53,55]
stats.ttest_rel(before_weights,after_weights)
```

Output:

```
Ttest_relResult(statistic=7.120275558034701,
pvalue=3.504936069662947e-06)
```

5. Conclusion/interpretation

Since the calculated p-value $< \alpha(0.05)$, we reject the null hypothesis.

It can be concluded that there is a significant difference between the two groups before and after the weight loss program.

Bibliography

<https://upload.wikimedia.org/wikipedia/commons/5/5b/Binomialverteilung2.png>

https://upload.wikimedia.org/wikipedia/commons/thumb/c/c1/Poisson_distribution_PMF.png/1200px-Poisson_distribution_PMF.png

https://upload.wikimedia.org/wikipedia/commons/thumb/f/fb/Normal_distribution_pdf.svg/900px-Normal_distribution_pdf.svg.png

https://commons.wikimedia.org/wiki/File:Standard_deviation_diagram.svg

https://upload.wikimedia.org/wikipedia/commons/d/d8/Normal_distribution_curve_with_lower_tail_shaded.jpg

https://upload.wikimedia.org/wikipedia/commons/thumb/c/cc/Relationship_between_mean_and_median_under_different_skewness.png/1200px

https://commons.wikimedia.org/wiki/File:Kurtosimailak_euskaraz.pdf

https://upload.wikimedia.org/wikipedia/commons/2/2d/Empirical_CLT_-_Figure_-_040711.jpg

https://upload.wikimedia.org/wikipedia/commons/1/10/Region_of_rejections_or_acceptance.png

Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Chi-square_distributionPDF.png)